Division of Population Health (DPH) - Data Classification Policy

This version (v2.0) Approved at IG Committee Meeting 2026-01-20

Background

The University of Sheffield has a central Information Classification & Handling Scheme which dictates how different types of information are to be stored. Non-research data within the Division will be held in accordance with the University policy. Research data, particularly data that carries a risk of identification (risk bearing data), will be classified and held in accordance with this policy, ensuring that security is proportionate to the sensitivity of identifiers.

Classification of data held within the Division of Population Health reflects the inherent risk of re-identification, i.e. it reflects the established understanding that achieving true anonymity for rich participant data is practically impossible. Even after removing direct identifiers, there is a risk that the unique combination of remaining details can be linked with external public information.

By classifying the data we hold we can ensure that security, access controls, and legal compliance (such as the maintenance of an Information Asset Register) are in place and proportionate to the risk (so that resources can be appropriately targeted).

Our policy is influenced by the 2019 Alan Turing Institute published paper called 'Design choices for productive, secure, data-intensive research at scale in the cloud' and aligns with current best practice for securing research data based on the level of identifying information it contains.

Scope

This policy applies to data held within the Division of Population Health (DPH) Departmental Storage (typically mapped as the "X: drive"). Data held within the Division of Population Health refers to data held within the X drive.
See also:

- Management of Division of Population Health (DPH) resources on Departmental Storage (typically mapped as the "X: drive")
- Routine maintenance of Division of Population Health resources on University Departmental Storage (typically mapped as the "X: drive")

Purpose

The purpose of this policy is to establish an Information Classification system for all data held within the Division of Population Health and help identify every project that holds Personal Data and/or Special Category Data.

The aim is to use this classification to ensure that data is stored and handled using technical and organisational safeguards appropriate to its sensitivity, which is mandatory under GDPR.

There is also a need to optimise resources, allowing Division of Population Health Data Security and Information Governance oversight to be targeted where required.

Currently the majority of research project data held by the Division of Population Health is subject to the same security requirements. This document is the starting point for introducing efficient, risk-based data protection.

There are no immediate actions required from researchers, rather this document aims to provide an introduction to some of the key principles of data classification and requirements around maintaining an information asset register.

Key principles and definitions

Information Classification:

This is the process of categorising data based on its sensitivity to determine the level of security required. Correct classification is necessary to apply the appropriate security measures.

Information Asset Register (IAR):

An IAR is a detailed, mandatory record of the organisation's or project's information assets: including all systems, files, and locations where personal data is stored or processed. The IAR must record: The type of data (e.g., special category), the location of the data (servers, secure drives), the Controller/Owner (responsible person), the security measures in place (e.g., encryption, access controls), the retention period (how long the data will be kept). Maintaining an IAR is a key part of the Accountability principle and helps demonstrate that the data is being stored securely.

Personal Data:

Personal data is any information that allows the identification of a person (the 'data subject'). This includes direct identifiers and indirect identifiers (see definitions below). If your data can be linked back to an individual, even through an intermediary code, it is personal data and falls under GDPR.

Special Category Data:

Special category data is personal data that is deemed highly sensitive and requires additional legal justification to process. For health research, the most relevant categories are data concerning health (physical or mental health), genetic data, and biometric data used for unique identification.

Commercial-in-Confidence Data:

Confidential business information that needs protection.

Pseudonymisation:

Pseudonymisation is the result of processing personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information (the Linkage Key). It is vital that the additional information (the Linkage Key) that would allow for the identification of individuals is kept separately and subject to strict security measures. **Pseudonymised data is still considered personal data under GDPR** but is viewed as a key safeguard for research.

Linkage ID

The "additional information" (as defined by GDPR) required to bridge the gap between the Pseudo ID and the original identity. Typically, a secure file where there is a column containing the Pseudo ID and columns containing the direct identifiers, e.g. name and contact details. It allows authorised users to "re-identify" the data if necessary (e.g., for follow-up).

Anonymisation:

Anonymisation is the result of processing data so that individuals can no longer be identified by any means reasonably likely to be used. Note that **in the context of rich, complex datasets like participant-level research data**, achieving true, irreversible anonymisation is very difficult due to the risk of re-identification through external data sources. The Information Commissioner's Office (ICO) has published guidance on anonymisation[1]. Data that is fully Anonymised is considered non-sensitive and falls outside the scope of the specific security controls required by this policy.

Data Minimisation:

Personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which it is processed. Researchers should only collect the data required for the study's aims and should pseudonymise or anonymise data at the earliest opportunity where the study purpose allows.

External Linkage Attacks:

Re-identification can occur if an attacker links an "anonymised" data to public information. Famous examples include:

[1] Anonymisation: managing data protection risk code of practice (ico.org.uk)

- Linking an "anonymised" Netflix movie rating dataset to publicly visible IMDb ratings to identify users.[2]
- Linking public voter registration lists (which contain name, address, gender, and birthdate) to "anonymised" health records to expose medical histories.[3]

The Context Problem:

In many research projects, the data itself contains descriptive details (a specific community, a unique job title, a quote about a specific event) that act as an indirect identifier. Since the researcher knows who the participants are, and the participants often know each other, the promise of anonymity is effectively broken within that context.

Direct identifiers

Name, address, full postcode, NHS number (or similar identifier), email address (if it contains other information about them e.g. name, date of birth), biometric data, facial photograph or comparable image / video, audiotapes (voice recordings), names of relatives, dates related to an individual (e.g. date of birth)

Direct identifiers with lower disclosure risk

Telephone number, email address (if no identifiable data included), initials, vehicle identifiers, medical device identifiers, internet protocol addresses

Indirect identifiers; may be identifiable if in combination or unusual

Place of treatment (e.g. hospital, site, GP surgery), partial postcode, health professional responsible for care (e.g. principal investigator, GP name), sex, rare disease or treatment, date of treatment / diagnosis / hospital visit, place of birth, socioeconomic data (e.g. rare occupation, place of work, income, or education), household and family composition, anthropometry measures, multiple pregnancies, ethnicity, year of birth or age (age is potentially identifying if the recruitment period is short and is fully described), verbatim (free text) responses or transcripts, religion. NB combining a participant's date of birth, gender, and general zip code has been found to uniquely identify a large proportion of people in large populations[4].

Unique Data Points (Quasi-Identifiers):

If your dataset contains rich research variables (e.g., a specific diagnosis, a rare genetic marker, or a specific behavioral score), there is a chance of reidentification.

Securely Separate Identifiers (separated)

---

[2] Narayanan, A. and Shmatikov, V., 2006. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.
[3] Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, *25*(2–3), 98–110.
[4] Sweeney, Latanya 2000, Simple Demographics Often Identify People Uniquely, 671 Health

This is the most common and stringent practice for achieving Pseudonymisation. It means isolating the direct identifying data (like names, addresses, or medical record numbers) from the research data and storing the two parts in different, secure locations, controlled by different access policies and/or different teams. The only link between them is a unique, non-meaningful code (the pseudonym). This separation is a key technical safeguard under the Privacy-by-Design approach.

Band (or Banding / Banded)

This refers to the technique of reducing the granularity or precision of data to make re-identification harder, thereby aiding Data Minimisation. Rather than storing an exact value, the data is grouped into a range or 'band.'

Data transformation

The set of techniques and processes applied to personal data to convert it into an anonymised (non-personal, non-identifiable), or pseudonymised form. The goal is to maximise data utility for analysis or sharing while minimising the privacy risk. There is also useful information regarding removing and transforming direct and indirect identifiers within Smith et al. (2015)[5] and Appendix 2 of the UKCRC guidance[6]

Introduction to Core Classification Criterion: Identifier Type

The classification level of research data is determined by the presence, type, and transformation status (i.e. degree to which identifying information has been altered (e.g. banded), removed, or separated) of identifiers it contains, reflecting its inherent risk of re-identification. It is assumed that this data is alongside special category (i.e. health) data, given the context is health research (Special Category Data).

Further details are in Appendix 1 (Detailed Classification Requirements)

Classification Level	Identifier Status (Highest Risk Pre	Risk Profile
A. Identifiable (Direct)	Contains Unredacted Direct Identifiers (e.g., Name, NHS Number).	High to Major Risk. Disclosure would likely cause moderate to major impact/harm to individuals or the University.

[5] Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials. Tudur Smith C, Hopkins C, Sydes M, Woolfall K, Clarke M, Murray G, Williamson P. April 2015.
[6] data_sharing_sop_guide_v1.1_-1.pdf (ukcrc-ctu.org.uk)

B. Linkable (Indirect)	Contains 2 or more Unredacted Indirect Identifiers (e.g., Partial Postcode, Date of Birth, Place of Treatment) that, when combined with other data (even public), could potentially identify an individual, OR contains a Linkage Key that could be linked to an individual.	Moderate to High Risk. Disclosure could cause minor to moderate impact/harm. Re-identification is possible but requires effort.
C. Participant Level/Restricted	Contains One or Zero Unredacted Indirect Identifiers (any Banded/Transformed Data is not considered an identifier) AND has No Linkage Key. OR Contains Non-PII data that requires standard restricted access (e.g., Commercial-in-Confidence data).	Low Risk. Disclosure would likely have little or no impact. Re-identification is highly unlikely.
D. Anonymised	Contains No Direct or Indirect Identifiers. Data is aggregated or statistically transformed to prevent re-identification. Out of Scope of this policy's security controls.	Minimal/Zero Risk. Data is not risk bearing.

Scope

This policy applies to all members of the Division who are responsible for managing data.

This policy is focused on risk bearing data, and is to help us sort our projects into clear categories based on how easily their data can be linked back to an individual person and if it contains any confidential business information. We are specifically looking at personal data, special category data and commercial-in-confidence data.

By clearly classifying the data, we aim to ensure we use the right level of protection for the right information, which helps us comply with the law and avoid spending time and money on security where it isn't needed.

The Division refers to risk bearing data, as defined in the Division of Population Health Information Governance Policy.

This policy will assign a Classification Level (Level A, B, or C) to risk bearing data based on its identifiers. Data that is fully Anonymised is considered non-sensitive and falls outside the scope of the specific security controls required by this policy.

Rationale for Policy: Identifier-Based Classification

The Division of Population Health has made the decision to base its research data classification scheme on a system based on the presence and nature of Direct and Indirect Identifiers; in order to make it more granular than the University Classification scheme. The reasons for this are outlined below.

1. Enhanced Precision and Clarity Based on Risk

- Identifiers Focus on Source of Risk: rather than a focus on the severity of the impact of a potential data breach (low, minor, moderate, major), which can be subjective and difficult to quantify during the initial project triage.
- A focus on the inherent identifying characteristics of the data itself (Direct vs. Indirect identifiers). This provides a clear, objective measure for classification that directly dictates the appropriate technical security controls. A data field's type (i.e. direct identifier), the combination of indirect identifiers, and the risk of identification through combination with another source determines its classification level, minimising ambiguity.

2. Alignment with UK and International Privacy Law

- The core concepts of data protection legislation (e.g., GDPR, ICO guidance) are built around the definition and handling of Personally Identifiable Information (PII) and the methods of pseudonymisation and anonymisation.
- By adopting an Identifier-based framework, our policy directly mirrors the language, principles, and risk assessment approach used by regulatory bodies. This makes the Division's data governance processes easier to audit and demonstrates a clear commitment to legal compliance.

3. Promoting Data Minimisation and Privacy-by-Design

- This structure creates a clear reminder for research teams to build into their project design the principles of data minimisation, i.e. remove, band, or securely separate identifiers. Though it is acknowledged this is already a key consideration at the ethics approval stage (refer also to https://staff.sheffield.ac.uk/rpi/ethics-integrity), this policy

should serve to illustrate the benefits. For example, by converting a Level B (Linkable/Indirect) dataset into a Level C (Pseudonymised/Restricted) dataset, researchers can significantly reduce their compliance burden and required security controls.
- This framework encourages a "Privacy-by-Design" approach, where risk reduction is embedded in the data collection and management phase, rather than being an administrative classification step applied afterwards.

In summary, shifting to an Identifier-based classification moves the Division from a subjective, impact-based system to an objective, risk-based system, leading to better compliance, clearer decision-making, and more scalable security management.

Process

Triage Process

When a request for a new data storage environment (e.g., a network folder creation request) is sent to Division of Population Health Data Security (DPH DS), the following information regarding the identifiers likely to be within data held in the project folder will be required to determine the appropriate Classification Level (A, B, or C) (see Appendix 2). This is an opportunity to consider data minimisation and privacy-by-design introduced in an earlier section (NB mid project reclassification is possible, see later section):

1. Direct Identifiers: Confirmation of whether the data held contains any Unredacted Direct Identifiers (e.g., Name, NHS number, Full Postcode) as defined in the Key principles and definitions section.
2. Indirect Identifiers & Pseudonymisation: If no Direct Identifiers are held by the study team, confirmation of whether the dataset contains 2 or more Unredacted Indirect Identifiers (see definitions) or if there is a Linkage Key that links the data back to an individual (even if not held with the rest of the data).

For projects where this essential identifier information is missing (e.g., folders created prior to this policy update), DPH-DS will gather this information as part of routine project reviews to ensure accurate classification.

The assigned Classification Level is determined by assessing the highest-risk identifier present (see also appendix 2):

- Level A (Identifiable): Assigned if the data held includes Unredacted Direct Identifiers.
- Level B (Linkable): Assigned if the data contains 2 or more Unredacted Indirect Identifiers or a Linkage Key.
- Level C (Participant Level/Restricted): Assigned if the data contains only Banded/Transformed Identifiers and there is no Linkage Key (NB if only one Unredacted Indirect Identifier that is considered Level C due to not being in combination with others).

Mid Project Re-Classification

There may be times during a project where the Classification Level needs reassessment. This change is typically driven by project activities that (potentially inadvertently) increase or reduce the data's risk of identification.

Triggers for Re-Classification

The need for re-classification is primarily triggered by actions that change the status of identifiers in the dataset, moving it to a higher risk level:

Classification Change	Triggering Action	Example
Level C → Level B	Exposure of Indirect Identifiers: The research process results in the accidental retention or creation of unredacted Indirect Identifiers (e.g., precise dates, fine-grained geographic data) within the main analysis file.	Combining several non-identifying datasets reveals highly specific data points that, when linked, can single out an individual.
Level C/B → Level A	Creation or Linkage to Direct Identifiers: Analysis or a protocol change requires the creation, collection, or introduction of Unredacted Direct Identifiers (e.g., obtaining an NHS number to perform a linkage step).	Researchers onboard known individuals onto the project and need to link data directly to them, or a linkage key is required to be held alongside the dataset.
Level A → Level C/B	Removal of Direct Identifiers: Participant follow up is complete and the Linkage ID is deleted.	Researchers complete participant follow-up and no longer require Direct Identifiers. The Linkage ID can be deleted and the pseudo ID no longer links to Direct Identifiers.

Process for Re-Classification

If any member of the research team suspects that the data's identifying status has changed, they must contact DPH-DS.

The process for managing a re-classification is as follows:

1. Notification: The Principal Investigator (PI) or a designated researcher must contact dph-DS to report the change in identifier status.
2. Risk Assessment: DPH-DS will confirm the highest-risk identifier present and assign the new Classification Level (A or B).
3. Security Uplift: The project environment must implement the necessary Security Controls required for the new, higher classification level. This may involve moving the data to a more secure environment.
4. Confirmation: Information Asset Owners (IAOs) will be required to confirm the new Classification Level as part of project reviews conducted by DPH-DS.

Appendix 1 Detailed Classification Level Requirements

Level A: Identifiable (Direct)

Requirement	Description
Security Control	Highest Level Required. Segregation of the direct identifiers from the analysis dataset is mandatory. Access restricted to strictly named individuals using multi-factor authentication. All activity must be auditable.
Use Cases	Projects requiring direct participant contact or linkage across non-pseudonymised registries.

Level B: Linkable (Indirect)

Requirement	Description
Security Control	Strong Controls. Segregation of the linkage key from the analysis dataset is mandatory. Access requires role-based control and multi-factor authentication. Analysis must often be performed only on 'slices' or partial datasets to manage risk.

Use Cases	Research requiring fine-grained geographical or temporal data, or projects where the keyholder is part of the research team but managed separately (e.g., holding the key in a highly restricted, separate folder).

Level C: Participant Level/Restricted

Requirement	Description
Security Control	Standard Restricted Controls. Access controls are applied based on data security, data handling, or access control requirements stipulated by the data provider.
Use Cases	General research analytics on de-identified datasets, commercial-in-confidence data providing a minor competitive advantage, or legally/politically sensitive data of trivial impact.

Appendix 2: Request form

This form will be embedded within the Management of DPH resources on Department Storage Process for new folders:

New folders

“Is the intention to store participant level data in this folder?”

If NO: The project is out of scope (Level D (Anonymised))

If YES: The request asks:

"Will the dataset contain any Unredacted Direct Identifiers?"

If YES: The project automatically requires Level A (Identifiable) classification, or the data management plan must be updated to include a process for immediate anonymisation/pseudonymisation which must be implemented.

If NO: The request asks:

"Will the dataset contain 2 or more Unredacted Indirect Identifiers or will there be a Linkage Key that links to individuals?"

If YES: The project requires Level B (Linkable) classification.

If NO: The project requires Level C (Participant Level/Restricted) classification (i.e. the data will contain one or zero Unredacted Indirect Identifiers and will not contain a Linkage Key).

Existing folders

In order to gather this information for existing folders the questions will be retrospective:

“Does the folder contain participant level data?”

If NO: The project is out of scope (Level D (Anonymised))

If YES: The request asks:

"Does the dataset contain any Unredacted Direct Identifiers (e.g. name, address, NHS number, date of birth) ?"

If YES: The project automatically requires Level A (Identifiable) classification.

(NB depending on the needs of your project consider if direct identifiers are needed. If not, update the data management plan to include a process for anonymisation/pseudonymisation: once this is implemented and all existing identifiable data in the folder destroyed the project can be re-classified.)

If NO: The request asks:

"Does the dataset contain 2 or more Unredacted Indirect Identifiers (e.g. ethnicity, age, place of treatment) or is there a Linkage Key that links to individuals data (e.g. at a hospital site)?"

If YES: The project requires Level B (Linkable) classification.

If NO: The project requires Level C (Participant Level/Restricted) classification (i.e. the data contains one or zero Unredacted Indirect Identifiers and will not contain a Linkage Key).

(NB We take a conservative, risk-based approach by assuming that participant level data remains vulnerable to future re-identification methods utilising external data sources, therefore security measures are still required.)

Version	Effective Date	Summary of changes
1.0	22/07/2024	n/a first version
2.0	20/01/2026	The Division of Population Health has made the decision to transition its research data classification scheme from a

		numeric 'Tier' system, as this ultimately wasn’t adopted due to it being difficult for researchers to interpret, to a system based on the presence and nature of Direct and Indirect Identifiers.